Computational Biology and High Performance Computing

Tutorial M4 a.m.

November 6, 2000
SC'2000, Dallas, Texas

Abstract

The pace of extraordinary advances in molecular biology has accelerated in the past decade due in large part to discoveries coming from genome projects on human and model organisms. The advances in the genome project so far, happening well ahead of schedule and under budget, have exceeded any dreams by its protagonists, let alone formal expectations. Biologists expect the next phase of the genome project to be even more startling in terms of dramatic breakthroughs in our understanding of human biology, the biology of health and of disease. Only today can biologists begin to envision the necessary experimental, computational and theoretical steps necessary to exploit genome sequence information for its medical impact, its contribution to biotechnology and economic competitiveness, and its ultimate contribution to environmental quality. High performance computing has become one of the critical enabling technologies, which will help to translate this vision of future advances in biology into reality. Biologists are increasingly becoming aware of the potential of high performance computing. The goal of this tutorial is to introduce the exciting new developments in computational biology and genomics to the high performance computing community.

Introduction

Horst Simon
HDSimon@lbl.gov
NERSC

Computational Biology and High Performance Computing

- **Presenters:**
  - Horst D. Simon
    - ✔ Director, NERSC
  - Manfred Zorn
    - ✔ Co-Head, Center of Bioinformatics and Computational Genomics, NERSC
  - Sylvia J. Spengler
    - ✔ Co-Head, Center of Bioinformatics and Computational Genomics, NERSC and Program Director, NSF
  - Craig Stewart
    - ✔ Director, Research & Academic Computing, Indiana University
  - Inna Dubchak
    - ✔ Staff Scientist, NERSC
- **Organizer:**
  - Manfred D. Zorn
- November 6, 2000

Tutorial Outline

- 8:30 a.m. - 12:00 p.m.
  - Introduction to Biology
  - Overview Computational Biology
  - DNA sequences
- 1:30 p.m. - 5:00 p.m.
  - Protein Sequences
  - Phylogeny
  - Specialized Databases

Tutorial Outline: Morning

- 8:30 a.m. - 8:45 a.m. Introduction
- 8:45 a.m. - 10:00 a.m. Biology
- 10:00 a.m. - 10:30 a.m. BREAK
- 10:30 a.m. - 12:00 p.m. Working with DNA

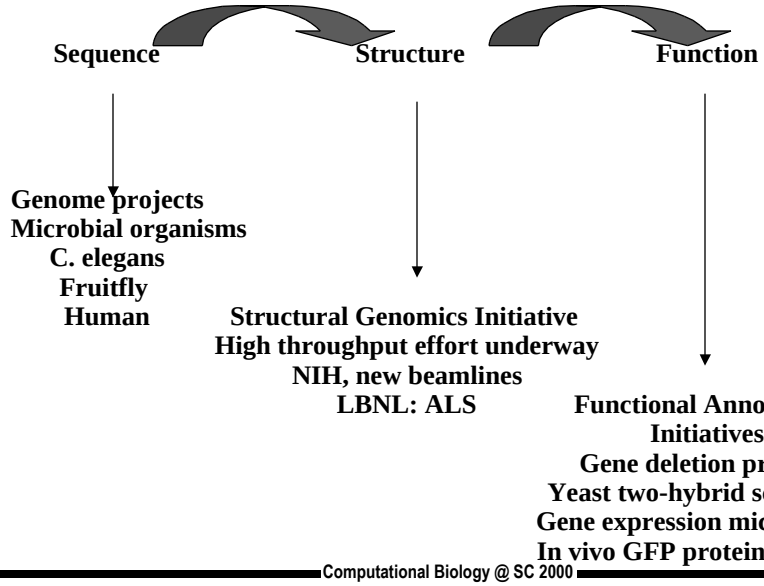
Tutorial Outline

- Introduction
- Brief Introduction into Biology
- DNA
  - What is DNA and how does it work?
  - What can you do with it?
- Proteins
  - What are proteins?
  - What do we need to know?
- Phylogeny
- Specialized Databases

Slide Credits

- Adam Arkin, LBNL
- Brian Shoichet, NorthWestern Univ.
- Teresa Head-Gordon, LBNL
- Sylvia J. Spengler, LBNL
- Manfred Zorn, LBNL
- Dodson-Hoagland: "The Way Life Works"
- National Museum of Health
  http://www.accessexcellence.org/
- B. Alberts et al. : "Essential Cell Biology"
  http://www.essentialcellbiology.com/
- L. Stryer: Biochemistry
- Genome Annotation Consortium
- Bob Robbins, FHCRC

Revolutionary Experimental Efforts in Biology



Computational Biology White Paper

http://cbcg.lbl.gov/ssi-csb

A technical document to define areas of biology exhibiting computational problems of scale

Organization:

- **Introduction to biological complexity and needs for advanced computing (1)**
- **Scientific areas (2-6)**
- **Computing hardware, software, CSET issues (7)**
- **Appendices**

For each scientific chapter:

- **illustrate with state of the art application (current generation hpc platform)**
- **define algorithmic kernals**
- **deficiencies of methodologies**
- **define what can be accomplished with 100 teraflop computing**

High-Throughput Genome Sequence Assembly, Modeling, and Annotation

The Genome Channel Browser to access and visualize current data flow, analysis and modeling. **(Manfred Zorn, NERSC)**

- **Genome sequencing and annotation ⟶ Bioinformatics**
  - **100,000 human genes; genes from other organism**
  - **Structure/functional annotation at the sequence level**
- **Computation to determine regions of a genome that might yield new folds**
  - **Experimental Structural Genomics Initiative**
  - **Functional annotation at the structure level by experiment**

Low Resolution Fold Topologies to High Resolution Structure

One microsecond simulation of a fragment of the protein, Villin. Duan & Kollman, Science 1998

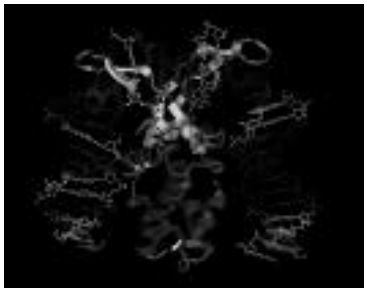
Low Resolution Structures from Predicted Fold Topology

- **Fold class gives some idea of biological function, but….**

Higher Resolution Structures with Biochemical Relevance

- **Drug design, bioremediation, diseases of new pathogen**

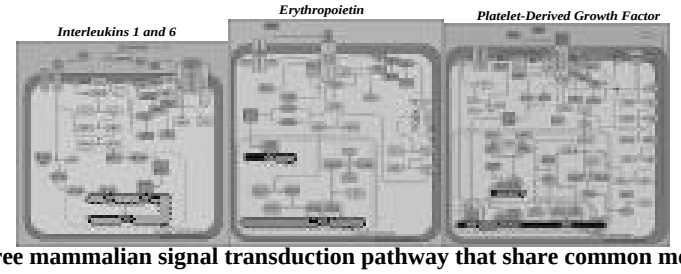
Simulating Molecular Recognition/Docking



Changes in the structure of DNA that can be induced by proteins. Through such mechanisms proteins regulate genes, repair DNA, and carry out other cellular functions.

- **Improvements in Methodology and Algorithms of Higher Resolution Structure**
  - **Breaking down size, time, lengthscale bottlenecks (IT^2, algorithms, teraflop computing)**
- **Protein, DNA recognition, binding affinity, mechanism with which drugs bind to proteins**
  - **Simulating two-hybrid yeast experiments**
  - **Protein-protein and Protein-nucleic acid docking**

Modeling the Cellular Program



Three mammalian signal transduction pathway that share common molecular elements (i.e. they cross-talk). From the Signaling PAthway Database (SPAD) (http://www.grt.kyushu-u.ac.jp/spad/)

- **Integrating Computational/Experimental Data at all levels**
  - **Sequence, structural functional annotation (Virtually all biological initiatives)**
- **Simulating biochemical/genetic networks to mode cellular decisions**
  - **Modeling of network connectivity (sets of reactions: proteins, small molecules, DNA)**
  - **Functional analysis of that network (kinetics of the interactions)**

The Need for Advanced Computing for Computational Biology

Computational Complexity arises from inherent factors:

- **100,000 gene products just from human; genes from many other organisms**
- **Experimental data is accumulating rapidly**
- **N^2, N^3, N^4, etc. interactions between gene products**
- **Combinatorial libraries of potential drugs/ligands**
- **New materials that elaborate on native gene products from many organisms**

Algorithmic Issues to make it tractable

- **Objective Functions**
- **Optimization**
- **Treatment of Long-ranged Interactions**
- **Overcoming Size and Time scale bottlenecks**
- **Statistics**

Introduction to Biology

Sylvia Spengler
SJSpengler@lbl.gov
NERSC

Biology

Cells

Proteins

DNA

DNA

Proteins

Cells

Scale

Comparative Scale of Mapping

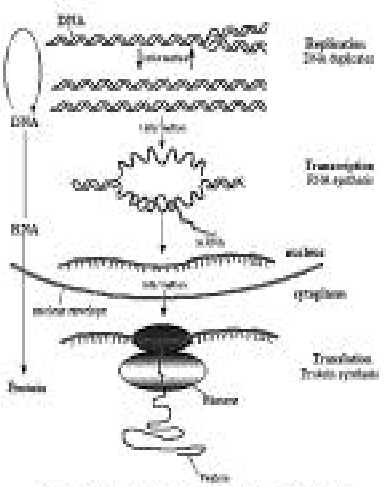
Truth and Conventional Wisdom in Biology

- Biologists dislike generalizations
- The truth in biology is always more complex than the statement about it
- It is hard to distinguish between fact and fashion in biology

Central Dogma

The fundamental dogma of molecular biology is that genes act to create phenotypes through a flow of information form DNA to RNA to proteins, to interactions among proteins (regulatory circuits and metabolic pathways), and ultimately to phenotypes.

Collections of individual phenotypes constitute a population.



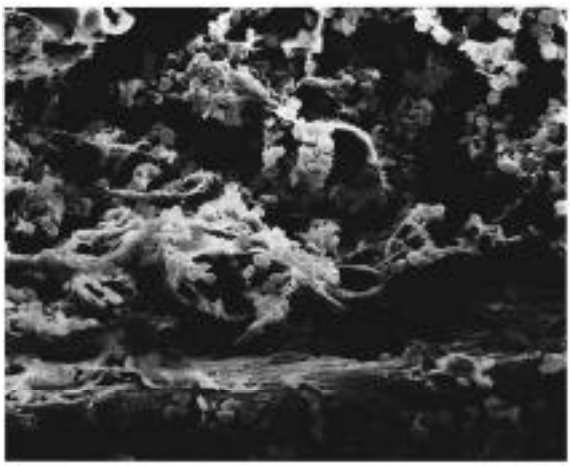
The Central Dogma of Molecular Biology

Biology is Special

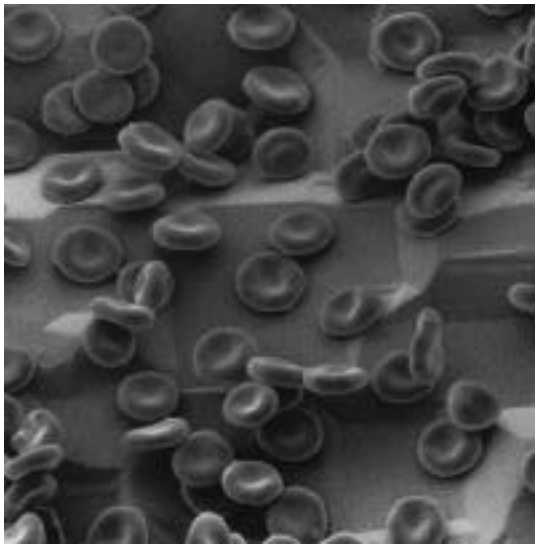
Life is characterized by

- Individuality
- Historicity
- Contingency
- high (digital) information content

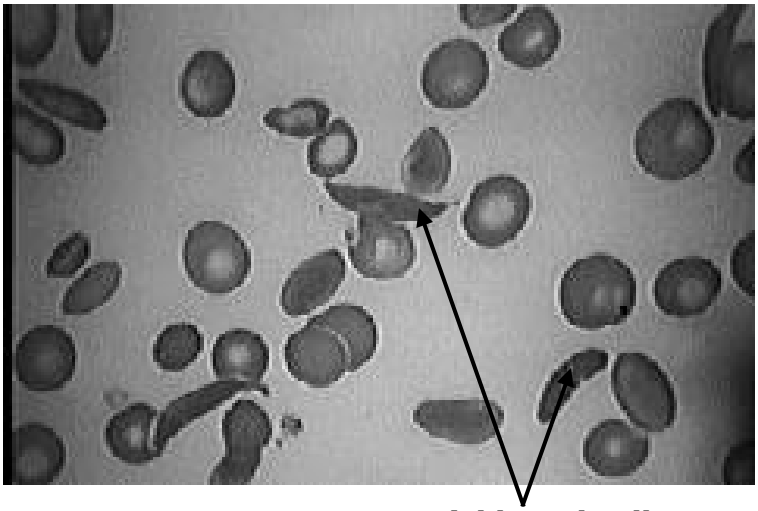
No law of large numbers, since every living thing is genuinely unique.



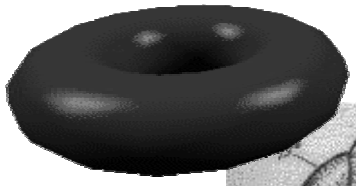
Chocolate Mints?

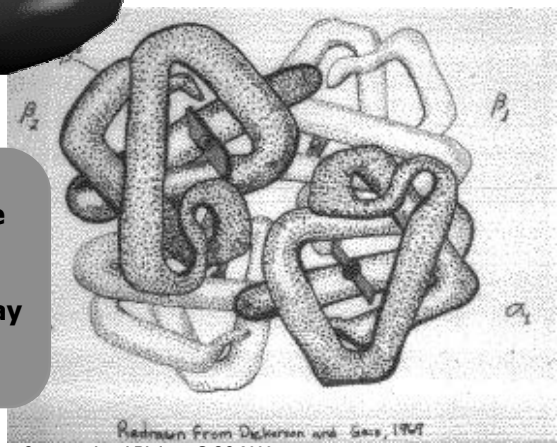


Diagnosis - Blood Smear



Red Blood Cells - Hemoglobin



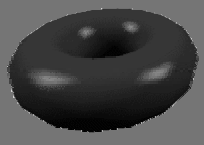


Hemoglobin is the main chemical in the red blood cell that does all of the work carrying oxygen away from the lungs and carbon dioxide back

Normal vs. Sickle Hemoglobin

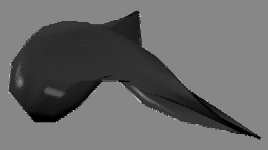
Normal

- disc-Shaped
- soft(like a bag of jelly)
- easily flow through small blood vessels
- lives for 120 days

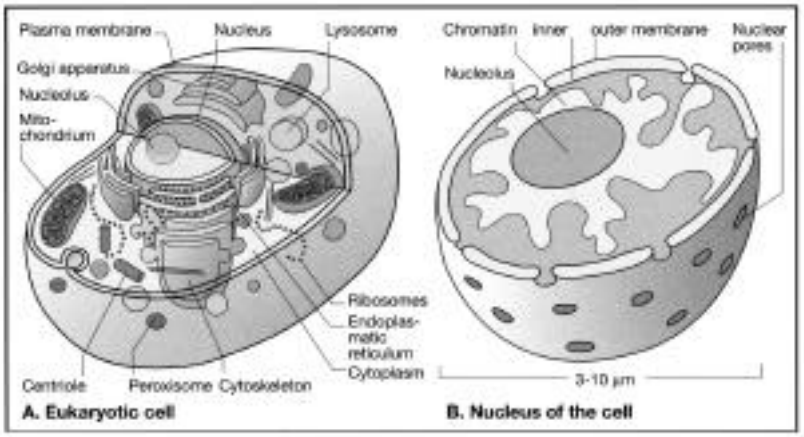


Sickle

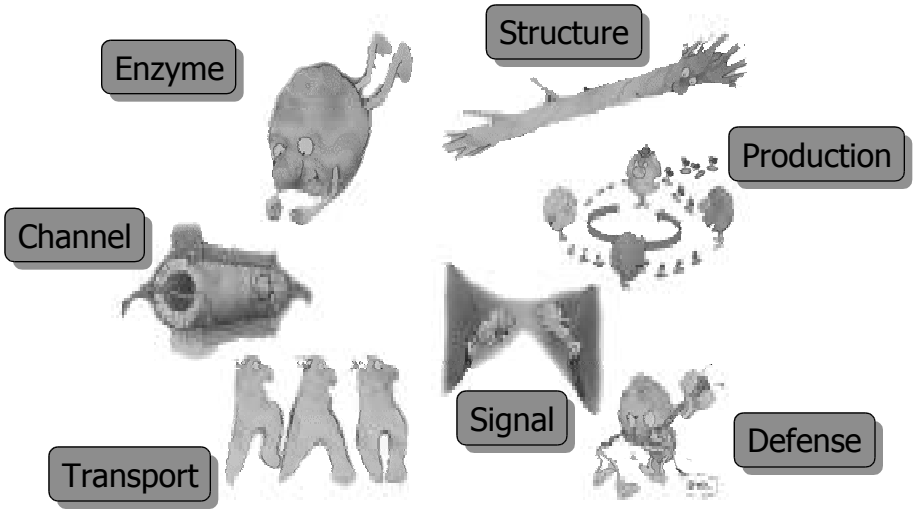
- sickle-Shaped
- hard (like a piece of wood)
- often get stuck in small blood vessels
- lives for 20 days or less



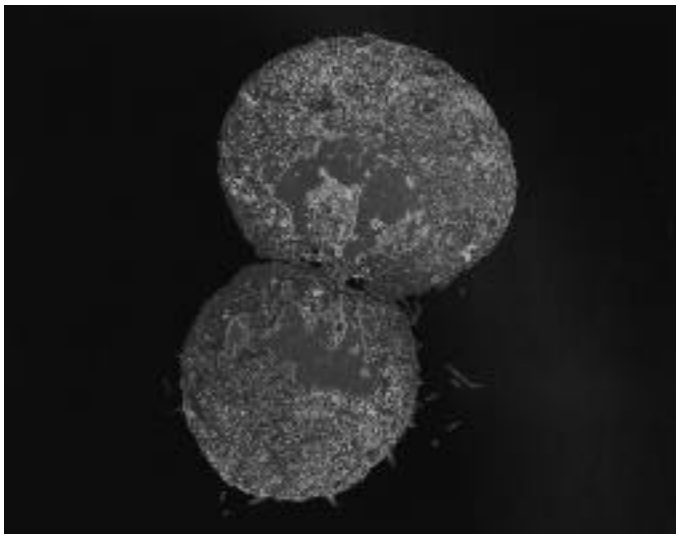
Cell Structure



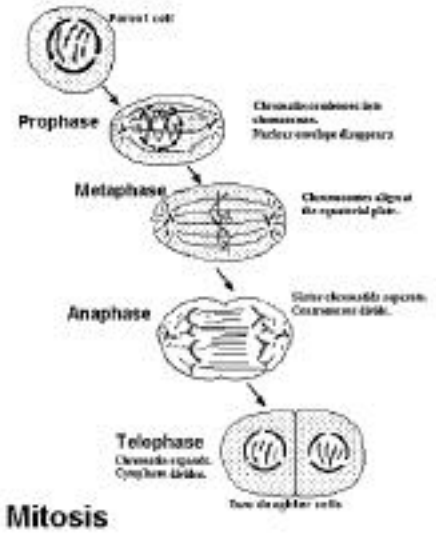
Protein Functions



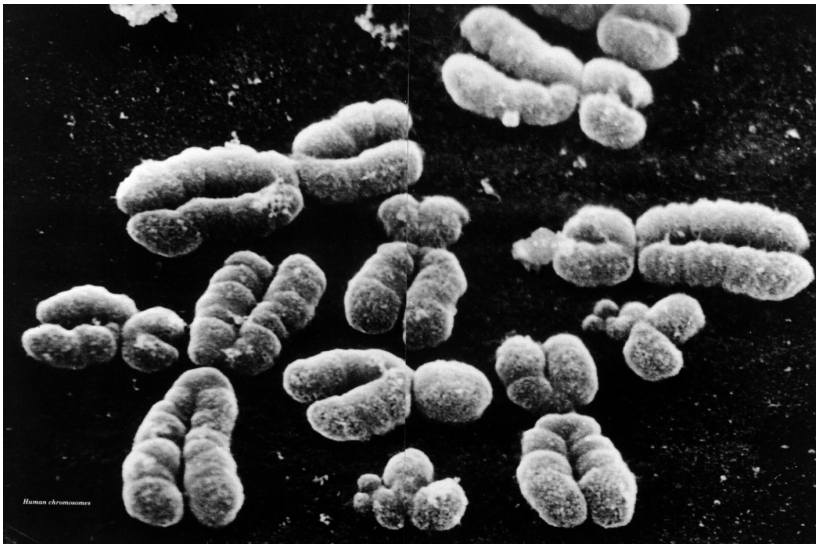
Cell Division



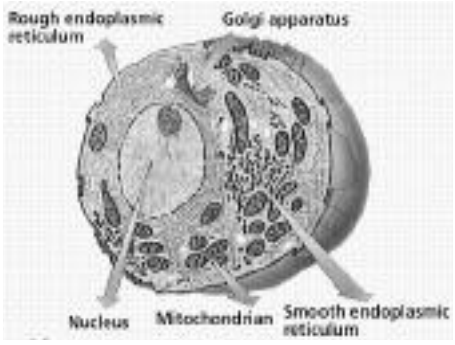
Cell Division



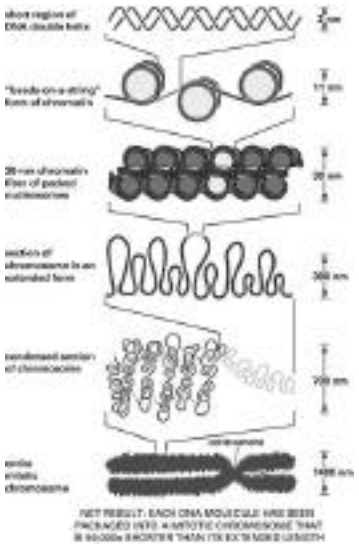
Chromosomes



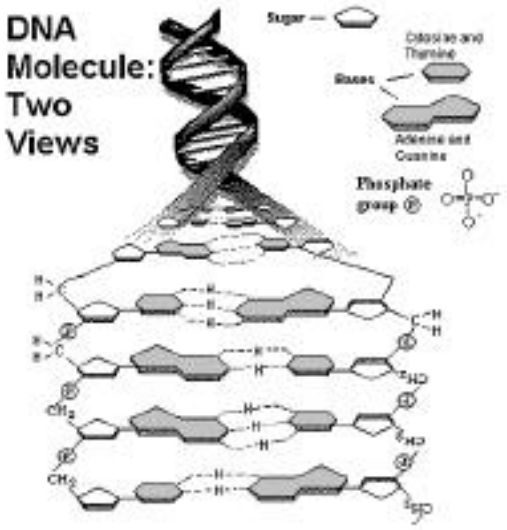
Basic Biology



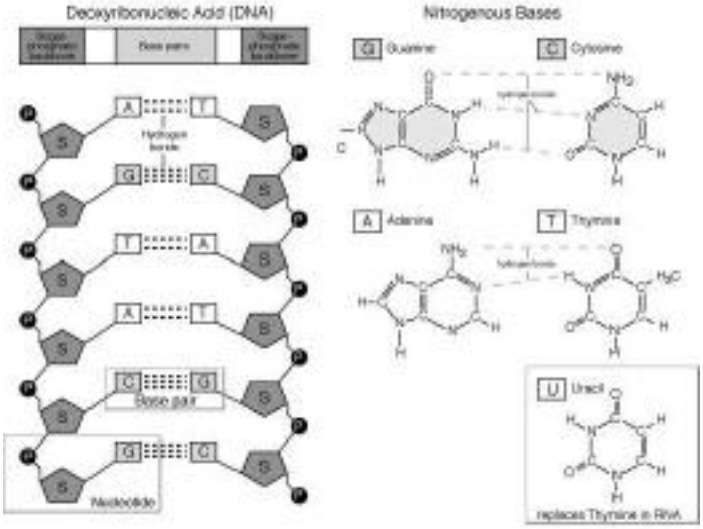
Scale



DNA - Two Views

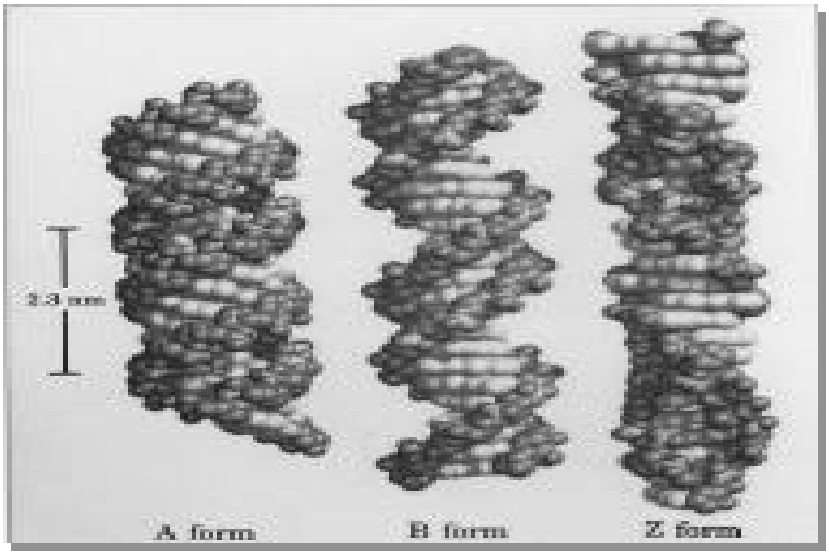


Four Bases

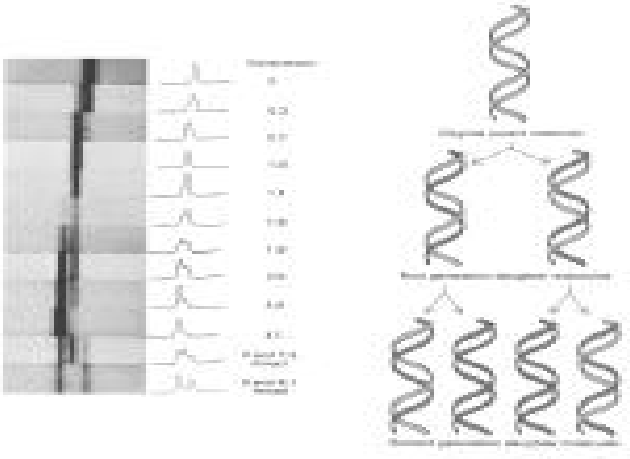


Double Helix

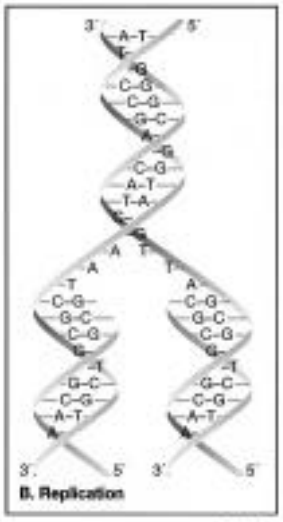
DNA



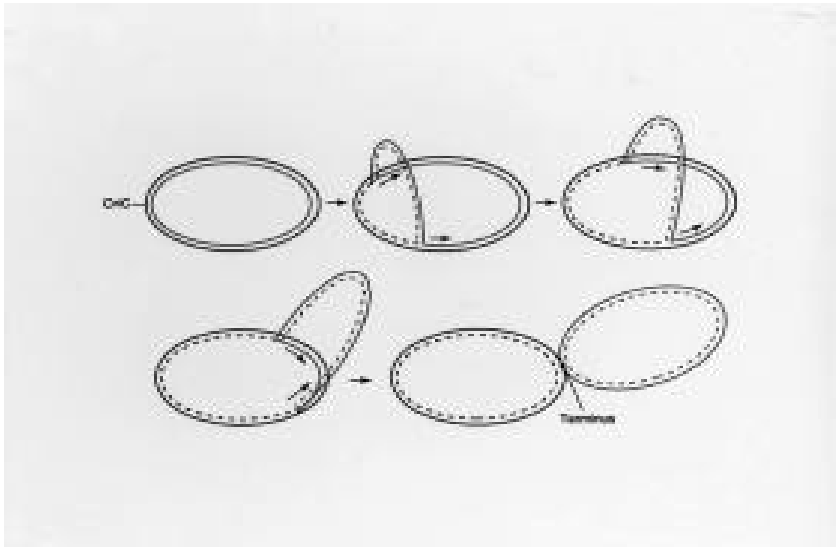
Semi-conservative Replication



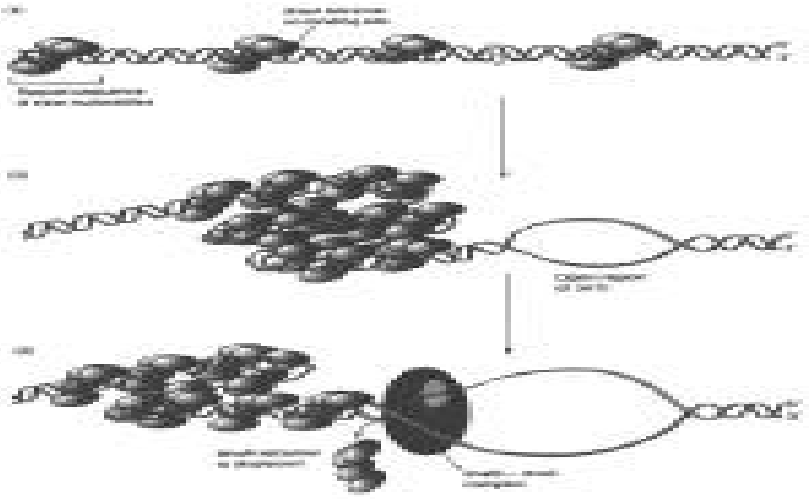
Replication



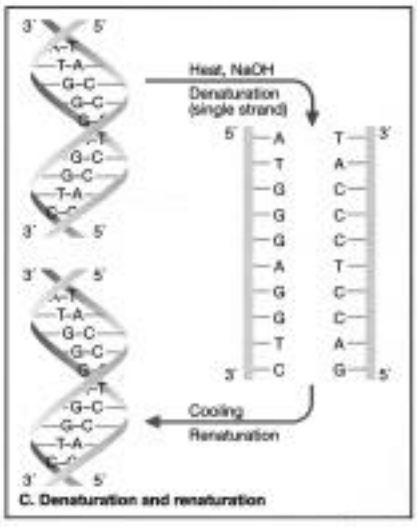
B. Replication



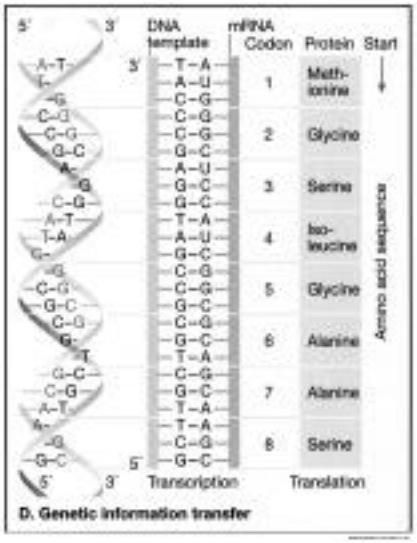
DNA Replication



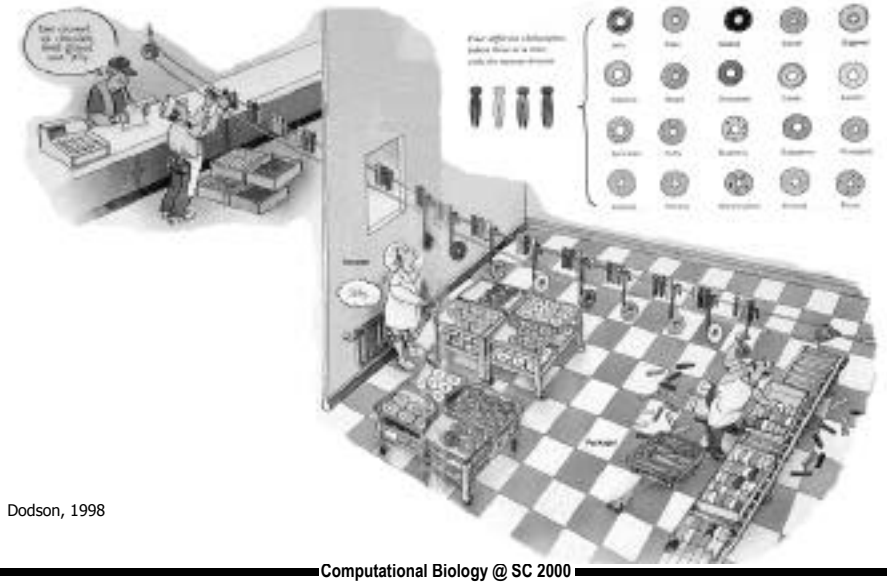
Hybridisation



Information Transfer

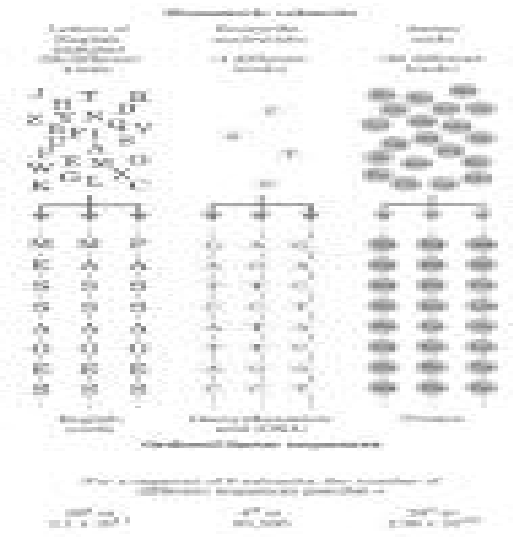


DNA Codes

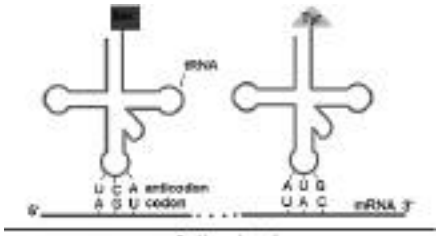


Dodson, 1998

Monomeric sub-units



Genetic Code

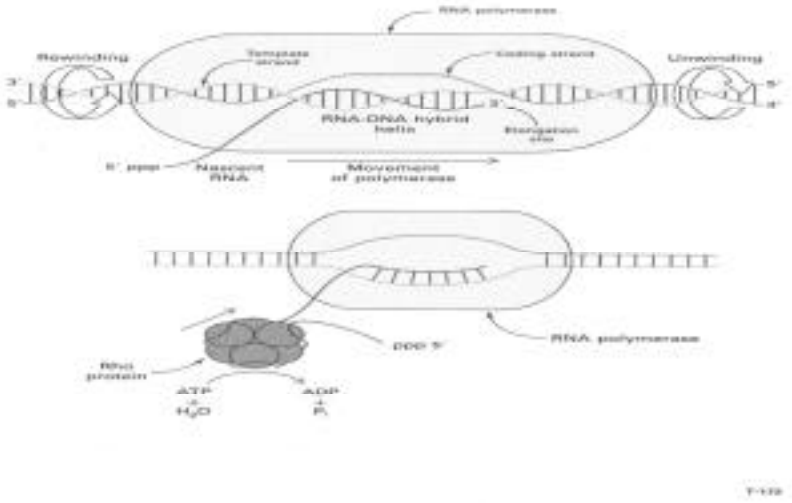


2nd base in codon

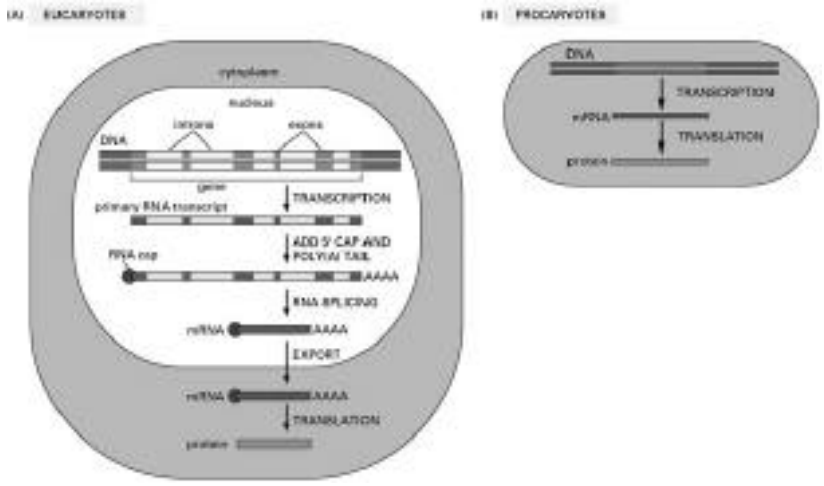
1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

The Genetic Code

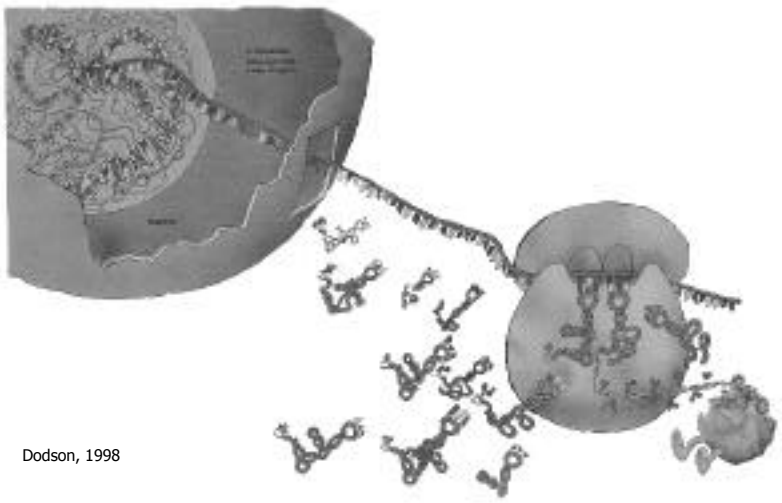
Transcription



Translation



Protein Construction



Dodson, 1998

Ribosome

Ribosome

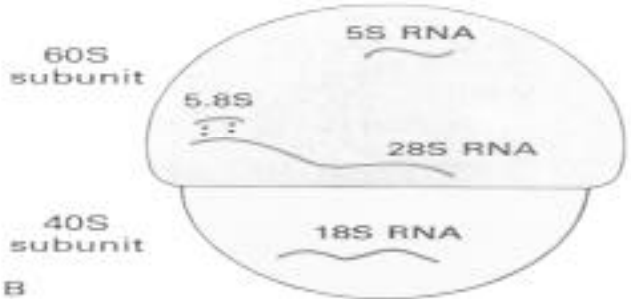


Figure 30-37
(A) Electron micrograph of eucaryotic ribosomes. [Courtesy of Dr. Miloslav Bublik.] (B) Schematic diagram of a eucaryotic ribosome.

RNA Base Pairs

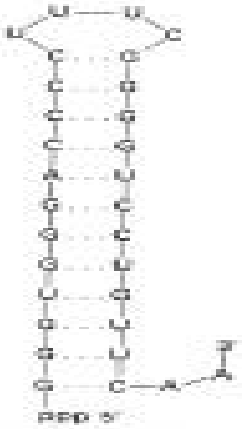


Figure 5-2
RNA can fold back on itself to form double-helical regions.

16S rRNA

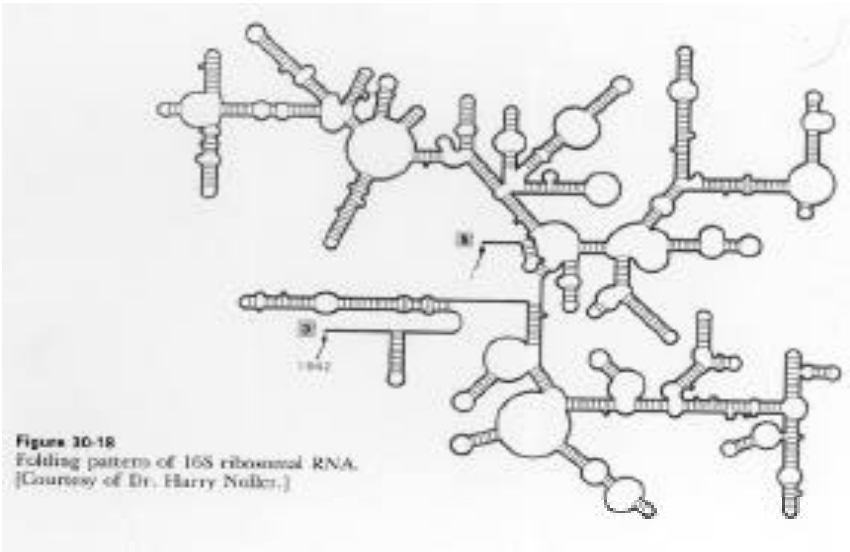
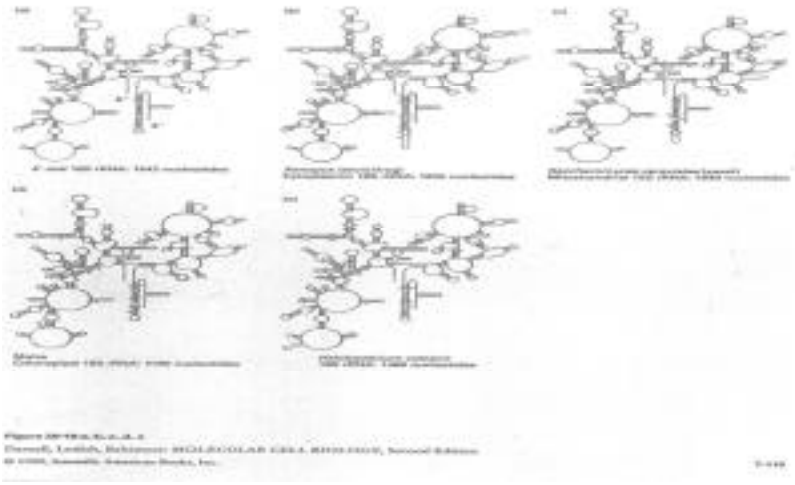
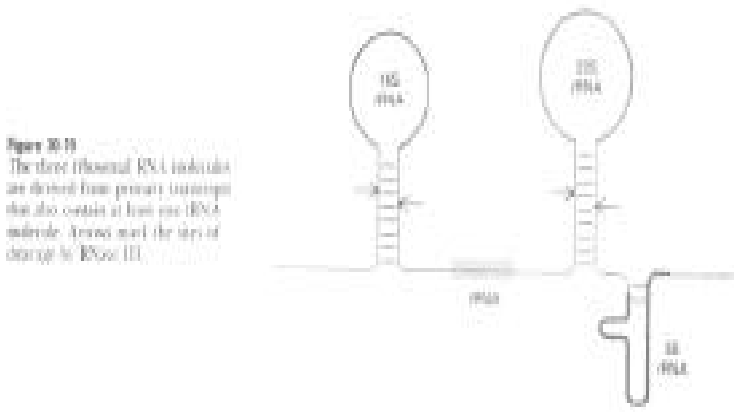


Figure 30-18
Folding pattern of 16S ribosomal RNA.
[Courtesy of Dr. Harry Noller.]

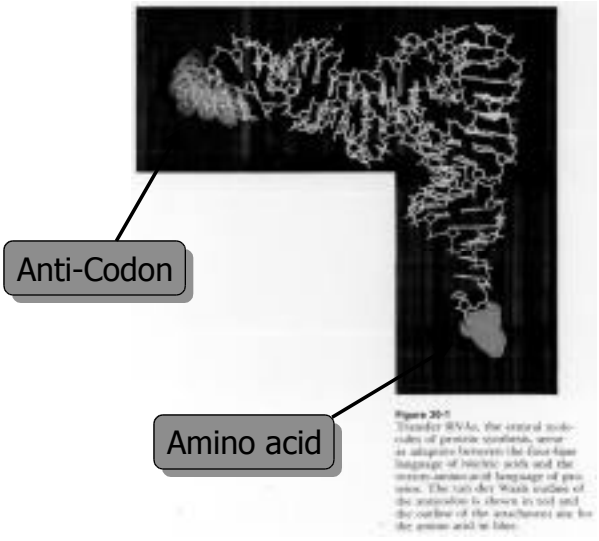
Small Subunit rRNA



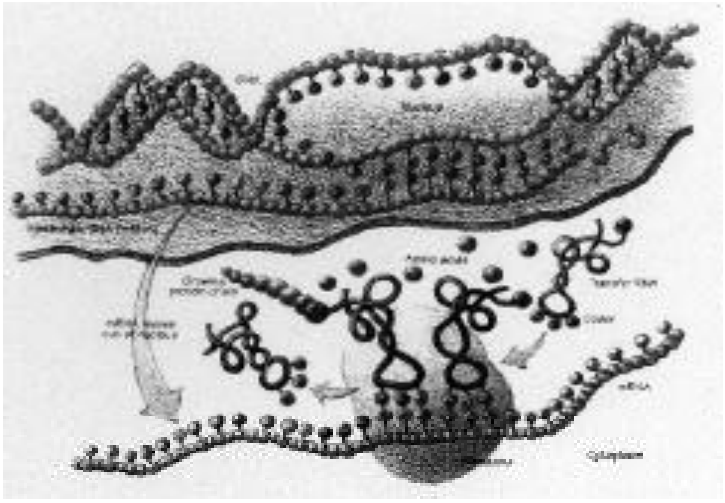
Cleavage by RNase III



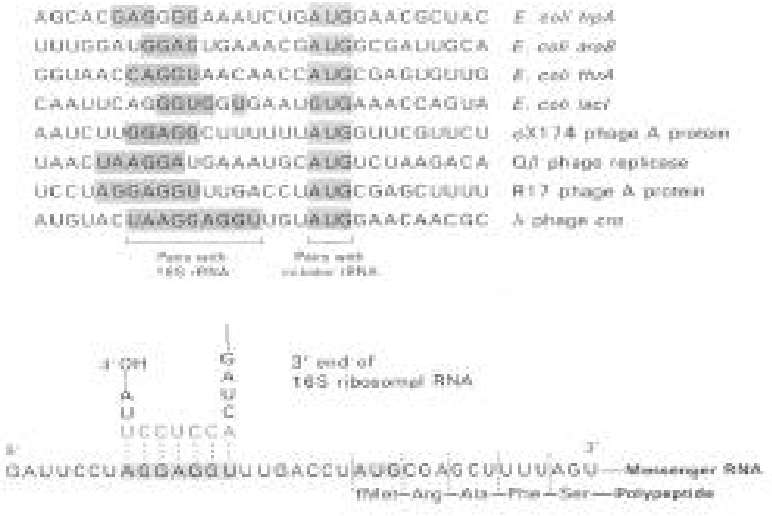
tRNA Structure



Protein Synthesis



Initiation

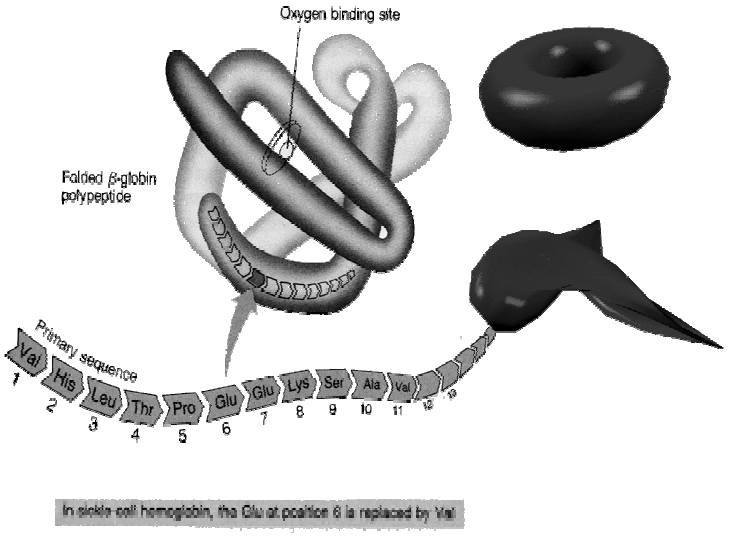


Sickle Mutation

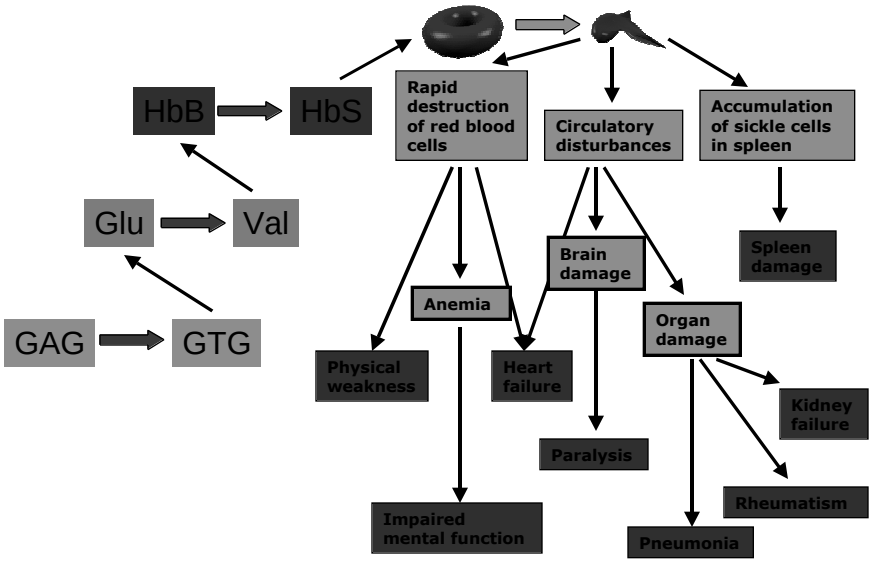
GAG (Glu)

↓

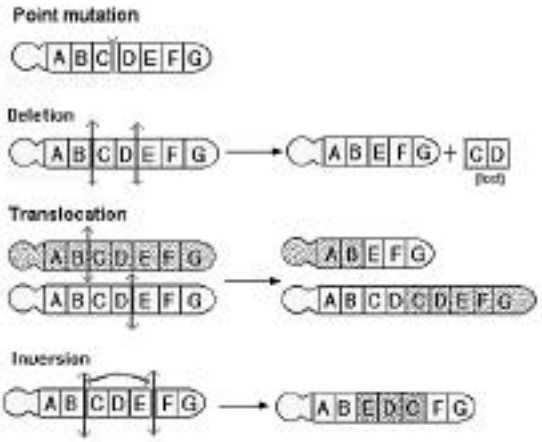
GTG (Val)



Sickle-cell Anemia

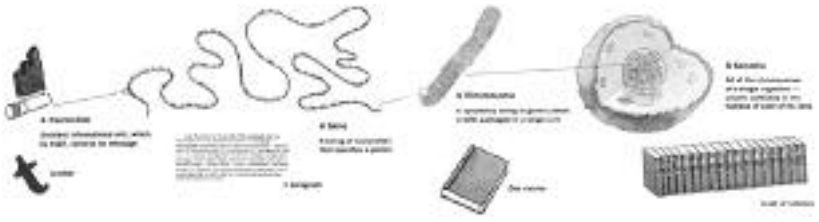


Mutations



Mutations of Chromosomes

Scale



Nucleomics

Manfred Zorn
MDZorn@lbl.gov
NERSC

Genome Project Timeline

- **1984**
  - ✔ **Department of Energy and Intl. Commission on Protection Against Environmental Mutagens and Carcinogens in Alta, Utah.**
- **1986**
  - ✔ **DOE announces Human Genome Initiative**
- **1987**
  - ✔ **NIH Director establishes Office of Genome Research**
- **1988**
  - ✔ **NRC Mapping and Sequencing the Human Genome**
  - ✔ **Berkeley Lab launches Human Genome Center**
- **1990 Human Genome I**

Genome Timeline cont'd

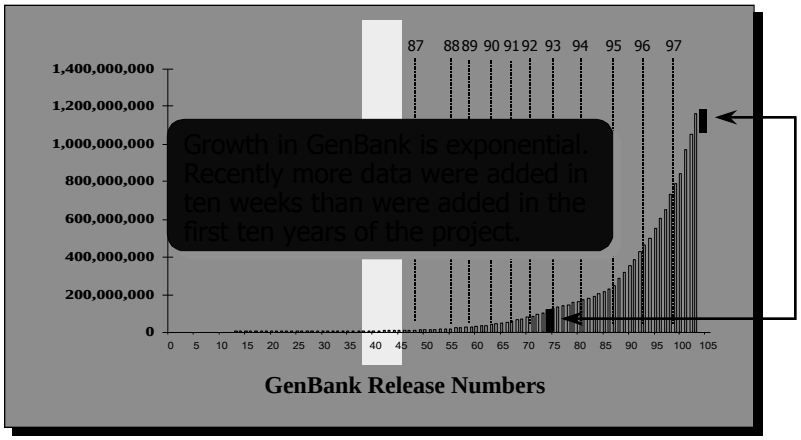
- **September 1994**
  - ✔ First complete map of all human chromosomes one year ahead of schedule.
- **May 1995**
  - ✔ First genome sequenced: H. influenzae
- **May 1998**
  - Celera announces commercial project
  - Public effort regroups to five major centers
- **June 2000**
  - Joint announcement by NIH/NHGRI - Celera

We're done!

Genome Projects

1995 H. influenzae	2 Mb
1996 S. cerevisiae	12 Mb
1997 E. coli	5 Mb
1998 C. elegans	100 Mb
1999 Human Chromosome 22	34 Mb
2000 D. melanogaster	140 Mb
2000 H. sapiens	3,000 Mb

Base Pairs in GenBank



DNA Sequencing

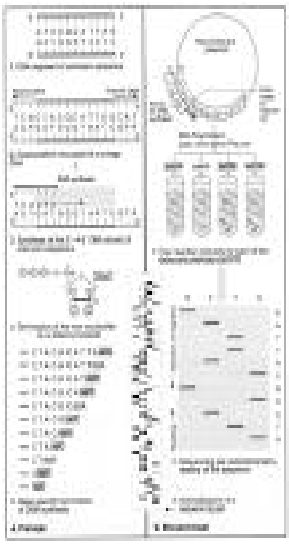
Read base code from storage medium!

- **Read length: About 600 bases at once**
- **Reader capacity**
  - ✔ **100 lanes in parallel in about 2-5 hours**
  - ✔ **1000 lanes in parallel in about 2 hours**

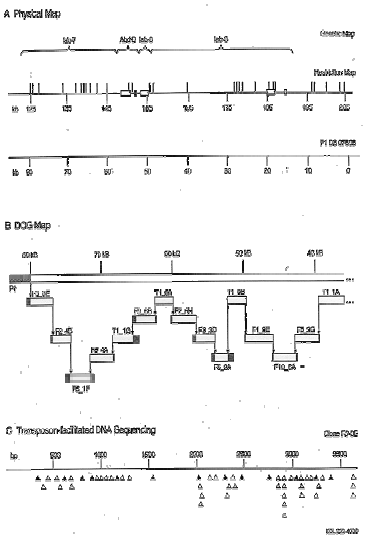
Sequencing: "bird's eye view"

- Prepare DNA
  - about a trillion DNA molecules
- Do the sequencing reactions
  - synthesize a new strand with terminators
- Separate fragments
  - by time, length = constant
- Sequence determination
  - automatic reading with laser detection systems

Sequencing



Mapping



Sequencing Strategies

Any genome is larger than amount of sequence that can be generated in a single step.

- **Shotgun**
- **Directed**
- **Finishing**

Shotgun

- **Break DNA into manageable pieces**
- **Sequence each piece**
- **Use sequence to reassemble original DNA**

Uniform process
Easily automatable

Coverage

$$\text{Coverage} = \frac{\text{Number x Size of clone}}{\text{Genome size}}$$

Expected gaps $\sim$ Number $e^{-\text{coverage}}$

Mapping project (Olson et al. 1986):
N=4,946
L=15,000
G=20,000,000
1,422 contigs vs. 1,457 predicted

Lander-Waterman 1988

Directed

- **Break DNA into manageable pieces**
- **Map pieces into tiling path**
- **Repeat**

Two separate processes: mapping and sequencing
More difficult to automate
Hard to integrate map information into assembly

- **T**

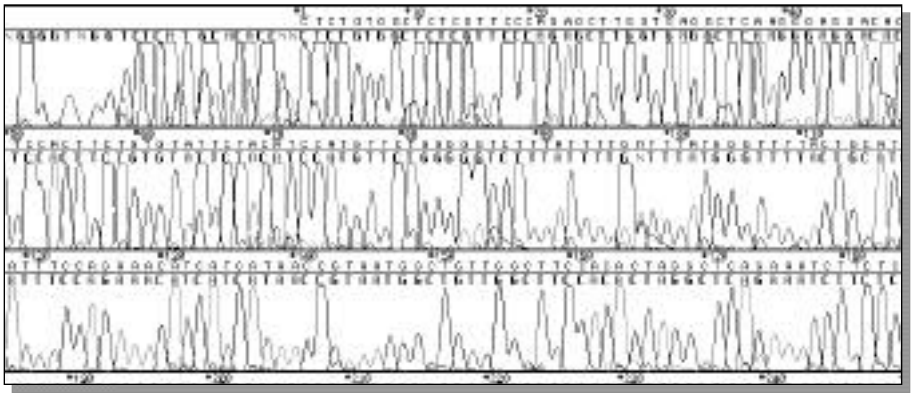
- **Use maps to assemble original DNA**

Finishing

- **Special cases that drop out of the pipeline**
- **Gap closing**
- **Difficult stretches**

- **Primer walking**
- **Different strains, vectors, chemistry**
- **Creative solutions, …….**

Sequence Traces



Good quality sequence needs about 10X Coverage

Base Calling

- **Machine records intensities in each channel**
- **Vendor software translates values into smooth signal for each base**
- **Base calling software "calls" the sequence**
- **Modern base callers use peak shape, size, and spacing as well as heuristics to improve quality of calls, i.e., fewer N's and better confidence.**
  - **Quality values carry base quality to the assembly step.**

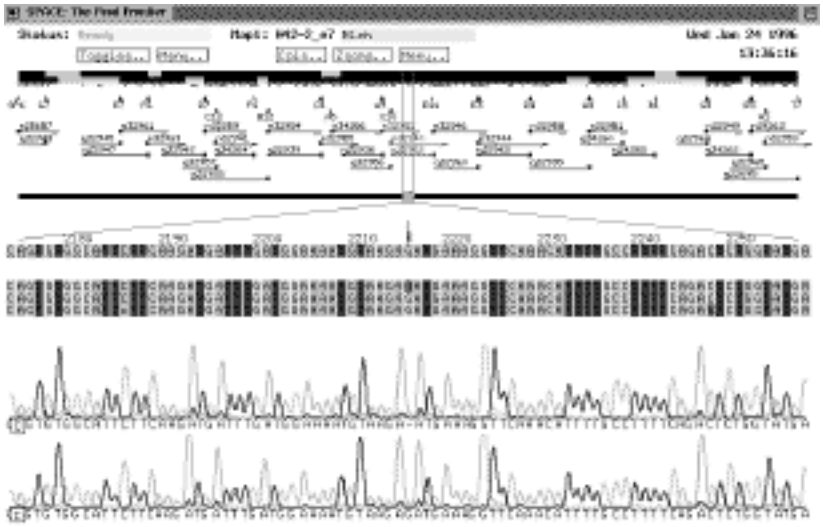
Phred - Base-caller

- Developed by Phil Green and Brent Ewing
- Better base calling accuracy
  - ✔ 40-50% lower error rates than ABI software on large test data sets
- Error probabilities for each base call
  - ✔ More accurate consensus sequences
  - ✔ Automatic identification of areas that require "finishing" efforts
  - ✔ Identification of repeat sequences in during assembly

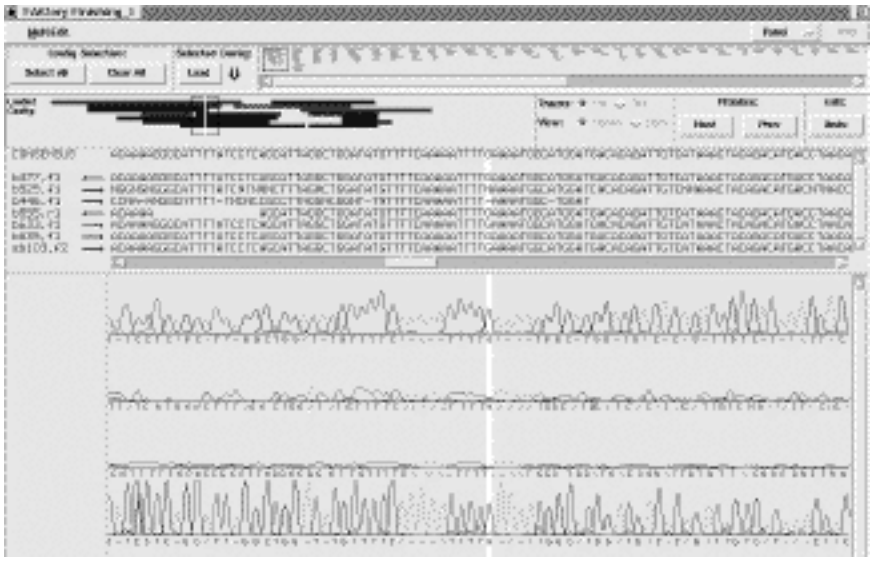
Phred's quality scores

After calling bases, Phred examines the peaks around each base call to assign a quality score to each base call. Quality scores range from 4 to about 60, with higher values corresponding to higher quality. The quality scores are logarithmically linked to error probabilities.

Quality score	Probability of wrong call	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%



FAKtory



Assembly

Putting humpty-dumpty together again!

- **Overlap**
  - ✔ **Find overlapping fragments**
- **Layout**
  - ✔ **Order and orientation of fragments**
- **Consensus**
  - ✔ **Determining the consensus sequence**
- **Use of constraints**

Assembly Features

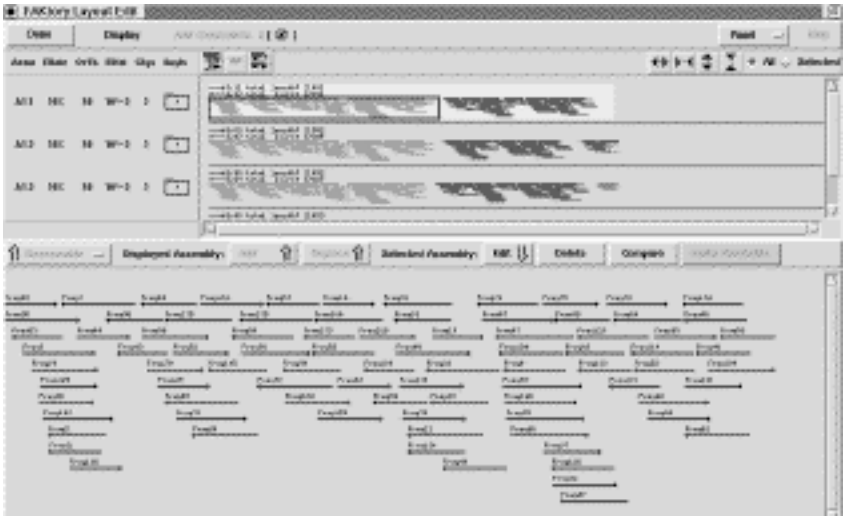
- **Repeats,**
  - **repeats,**
    - ✔ **repeats,**
      - **Repeats**
      - **200 bp Alu repeat every ~4,000 bp with 5% -15% error**

- **Clipping**
- **Orientation**
- **Contamination**
- **Rearrangements**
- **Sequencing errors**
- **True Polymorphisms**

Phrap - Assembler

- **Fast assemblies**
  - ✔ **Projects with several hundred to two thousand reads typically take only minutes**
- **Accurate consensus sequences from mosaic**
  - ✔ **Examines all individual sequences at a given position, and generally uses the highest quality sequence to build the consensus.**
- **Consensus quality estimates**
  - ✔ **Quality information of individual sequences yields the quality of the consensus sequence**
  - ✔ **Other available information about sequencing chemistry (dye terminator or dye primer) and confirmation by "other strand" reads used in estimating the consensus quality.**

FAKtory Layout



More assembly

- **Finishing: closing gaps**
- **Building chromosomes from large contigs that are consistent with map information**

What is a Gene?

- **Definition: An inheritable trait associated with a region of DNA that codes for a polypeptide chain or specifies an RNA molecule which in turn have an influence on some characteristic phenotype of the organism.**

Abstract concept that describes a complex phenomenon

What is Annotation?

- **Definition: Extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge.**

Identifiable features in the sequence

How does an annotation differ from a gene?

- **Many annotations describe features that constitute a gene.**
- **Other annotations may not always directly correspond in this way, e.g., an STS, or sequence overlap**

DNA Analysis

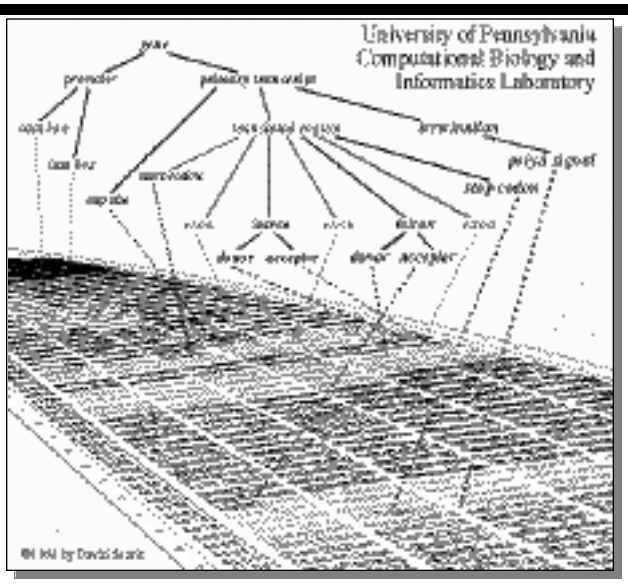
- Heuristics
- Statistics
- Artistics

DNA Analysis

Disassemble the base code!

- Find the genes
  - Heuristic signals
  - Inherent features
  - Intelligent methods
- Characterize each gene
  - Compare with other genes
  - Find functional components
  - Predict features

What is a Gene?

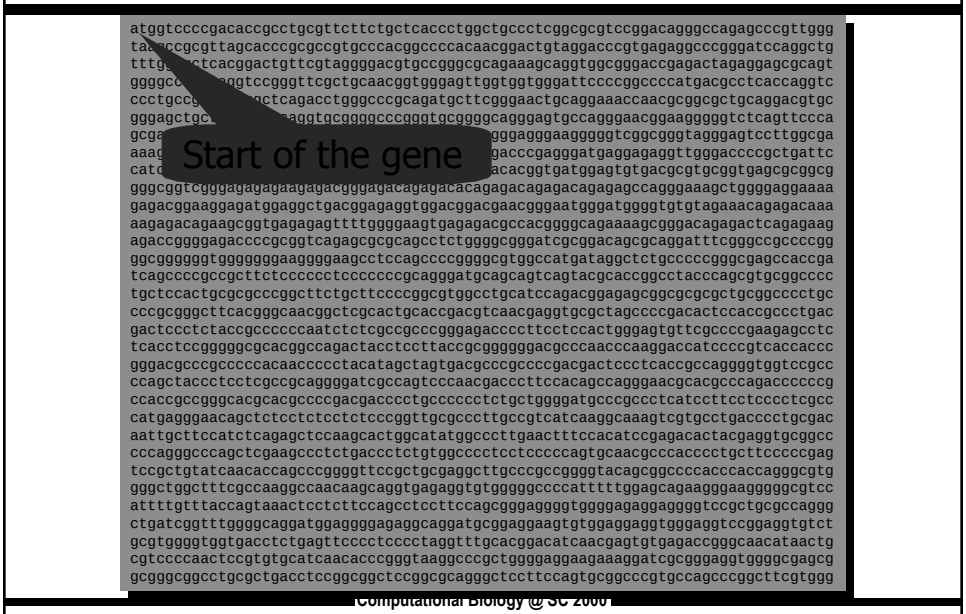


Heuristic Signals

DNA contains various recognition sites for internal machinery

- **Promoter signals**
- **Transcription start signals**
- **Start Codon**
- **Exon, Intron boundaries**
- **Transcription termination signals**

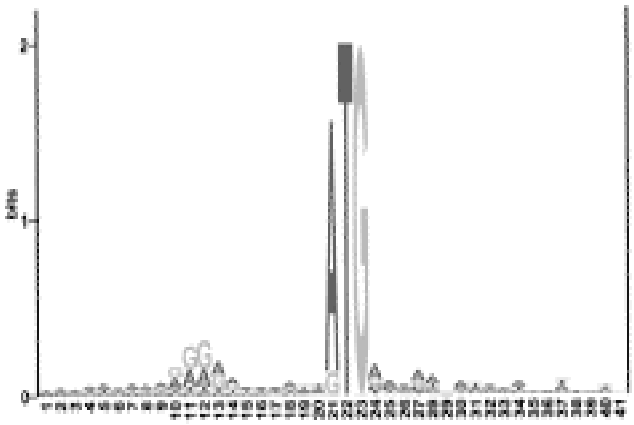
Heuristic Signals



Heuristic Signals

Computational Biology @ SC 2000

Start Codon

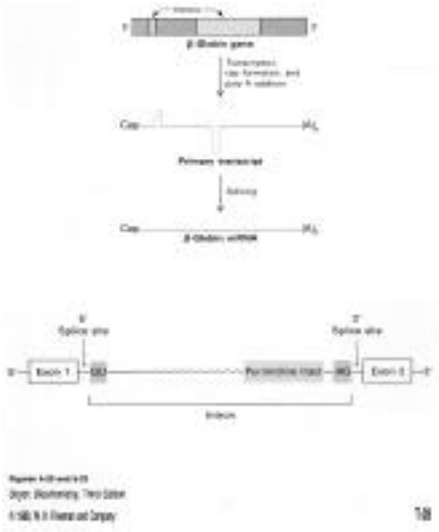


Inherent Features

DNA exhibits certain biases that can be exploited to locate coding regions

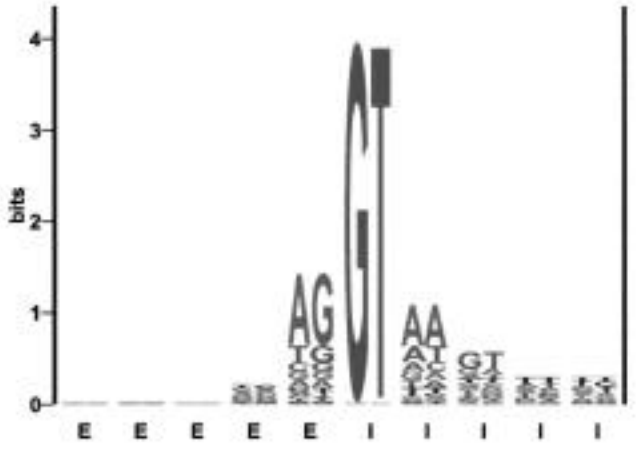
- **Uneven distribution of bases**
- **Codon bias**
- **CpG islands**
- **In-phase words**
- **Encoded amino acid sequence**
- **Imperfect periodicity**
- **Other global patterns**

Splicing

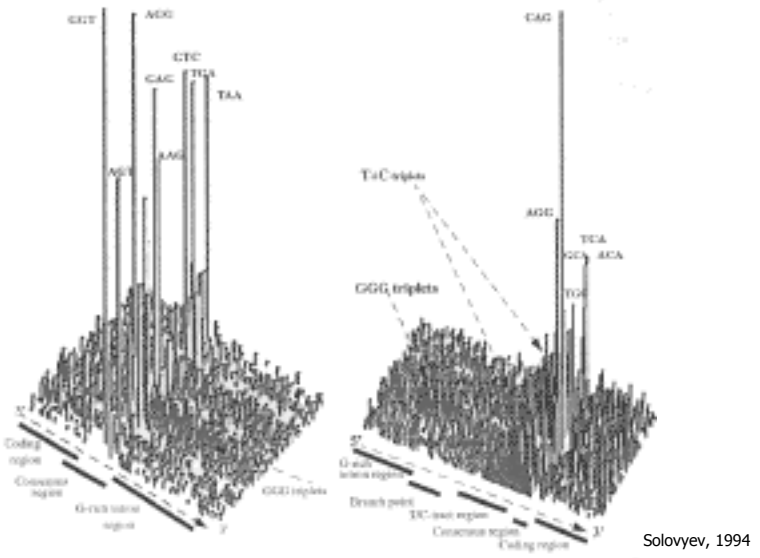


Donor Splice Site

Plate IV: A Logo of Donor Splice Sites from the Dicot Plant *A. thaliana* (cress). See page 34 for full discussion.



Inherent Features



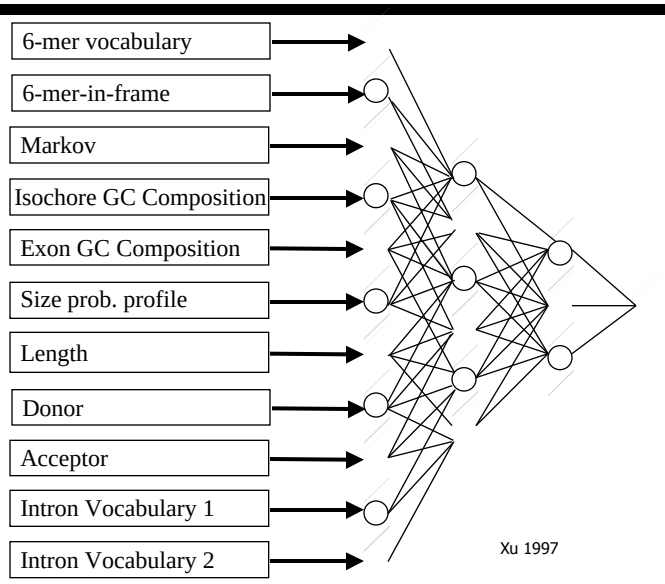
Solovyev, 1994

Intelligent Methods

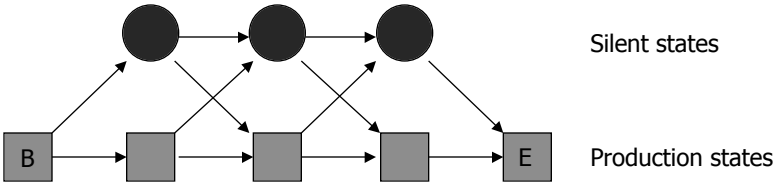
Pattern recognition methods weigh inputs and predict gene location

- **Neural Networks**
- **Hidden Markov Models**
- **Stochastic Context-Free Grammer**

Neural networks



Hidden Markov Models



Characterize a Gene

Collect clues for potential function

- Comparison with other known genes, proteins
- Predict secondary structure
- Fold classification

- Gene Expression
- Gene Regulatory Networks
- Phylogenetic comparisons
- Metabolic pathways

Comparison with other sequences

- Dynamic programming
  - Needleman - Wunsch
  - Smith - Waterman
  - Evolution

- Speed vs. sensitivity
  - Hashing
  - Statistical considerations
  - Suffix trees

Terminology

- **Homology**
  - ✔ **Common ancestry**
  - ✔ **Sequence (and usually structure) conservation**
  - ✔ **Homology is not a measurable quantity, but can be inferred, under suitable conditions**
- **Identity**
  - ✔ **Objective and well defined**
  - ✔ **Can be quantified by several methods:**
    - ✦ **Percent**
    - ✦ **The number of identical matches divided by the length of the aligned region**
- **Similarity**
  - ✔ **Most common method used**
  - ✔ **Not so well defined**
  - ✔ **Depends on the parameters used (alphabet, scoring matrix, etc.)**

Alignment

- **An alignment is an arrangement of two sequences opposite one another**
- **It shows where they are different and where they are similar.**

We want to find the optimal alignment - the most similarity and the least differences

Alignment

- Alignments have two aspects:
  - Quantity: To what degree are the sequences similar (percentage, other scoring method)
  - Quality: Regions of similarity in a given sequence

How is an alignment done?

- When we compare sequences, we take two strings of letters (nucleotides or amino acids) and align them.
- Where the characters are identical, we give them a positive score, and where they differ, a negative value.
- We count the identical and nonidentical characters, and give the alignment a score (usually called the quality)

Dynamic Programming

- **Sequence A**
- **Sequence B**
- **Substitution**
- **Deletion**
- **Insertion**
- **Matrix Element**

$$A = (A_1, \ldots A_m)$$

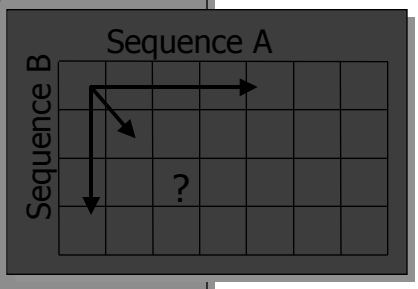
$$B = (B_1, \ldots B_n)$$

$$\omega(A_i, B_j)$$

$$\omega(A_i, \Delta)$$

$$\omega(\Delta, B_j)$$

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + \omega_{A_i, B_j} \\ H_{i,j-1} + \omega_{A_i, \Delta} \\ H_{i-1,j} + \omega_{\Delta, B_j} \end{cases}$$



Differences in the sequence can be caused by deletions or insertions in the DNA, or by point mutations. These changes can be seen at the protein level as well (changes in the translation of the protein

This scheme works fine as long as you assume that all possible mutations occur at the same frequency. However, nature doesn't work this way. It has been found that in DNA, transitions occur more often than transversions.

Scoring Matrices

- Identity scoring
- Genetic code scoring
- Physical chemical similarities
- Observed substitutions
  - Dayhoff matrix (PAM)
  - BLOSUM

The Gap Penalty

Consider the two following alignments:

```
V I T K L G T C V G S          V I T K L G T C V G S
V I T . . . T C V G S          V . T K . G T C V . S
```

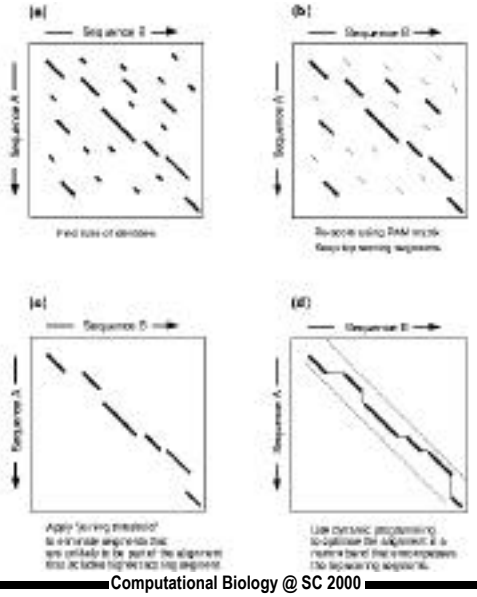
According to the algorithm these two cases will get the same gap penalty. However, in nature in most cases insertions/deletions are longer than just a single residue, even for very homologous sequences.

- To compensate for this, and to differentiate between cases like the one above, the gap penalty is made up of two factors:
  - The gap creation penalty - subtracted from the alignment quality whenever a gap is opened.
  - The gap extension penalty - subtracted from the alignment quality according to the length of the gap.

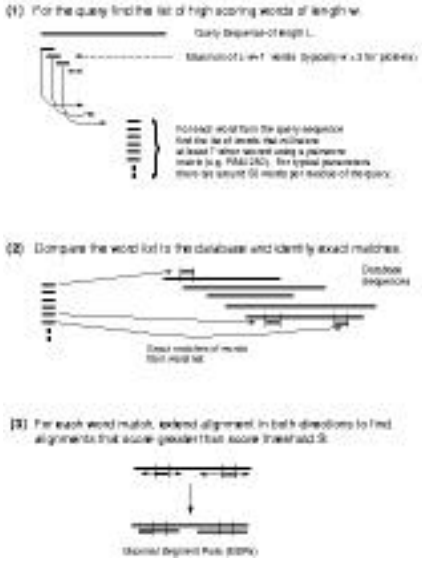
Score

- Thus we have the following score:
  - Quality = matches - (mismatches + gap penalty)
  - Gap penalty = gap creation penalty + (gap extension penalty X gap length)

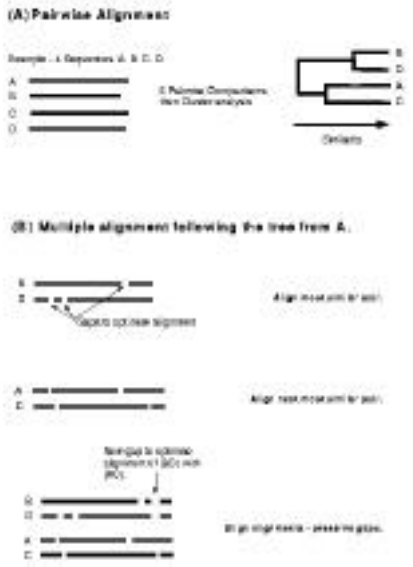
FASTA



BLAST

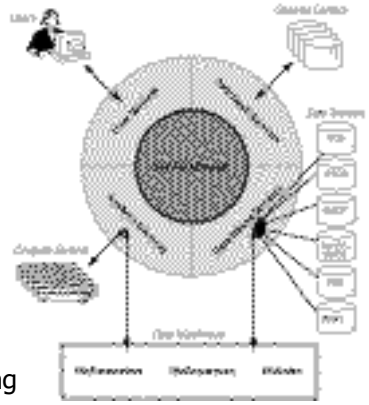


Multiple Alignments

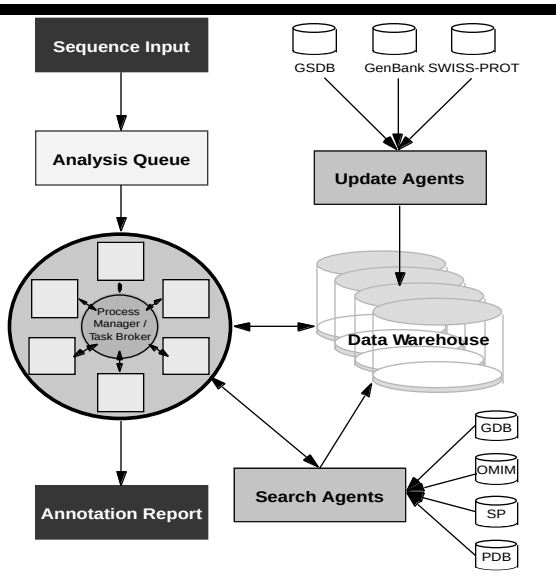


Large-scale Genome Annotation

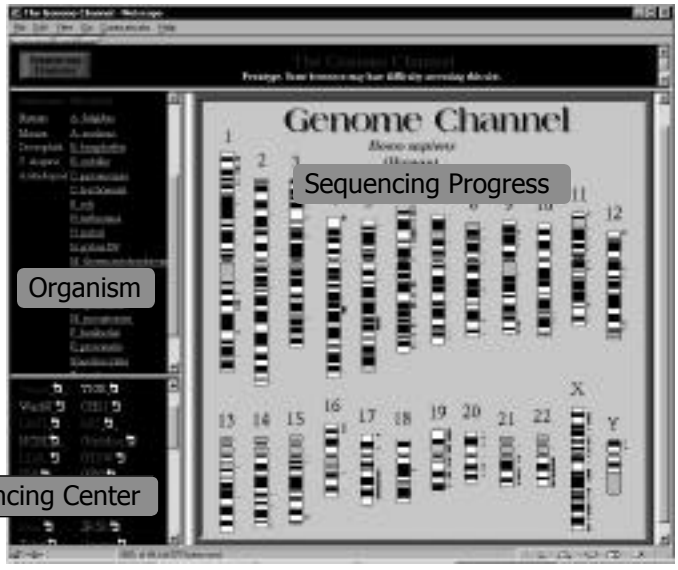
- Multi-laboratory Project
- Standard Annotation of Genomes
  - Genome Channel
  - Genome Catalog
- Comprehensive integration of
  - Analysis tools
  - Data management systems
  - Data mining
  - User services
- Extensible Framework
  - High-performance computing
  - Data integration technology
  - Artificial intelligence



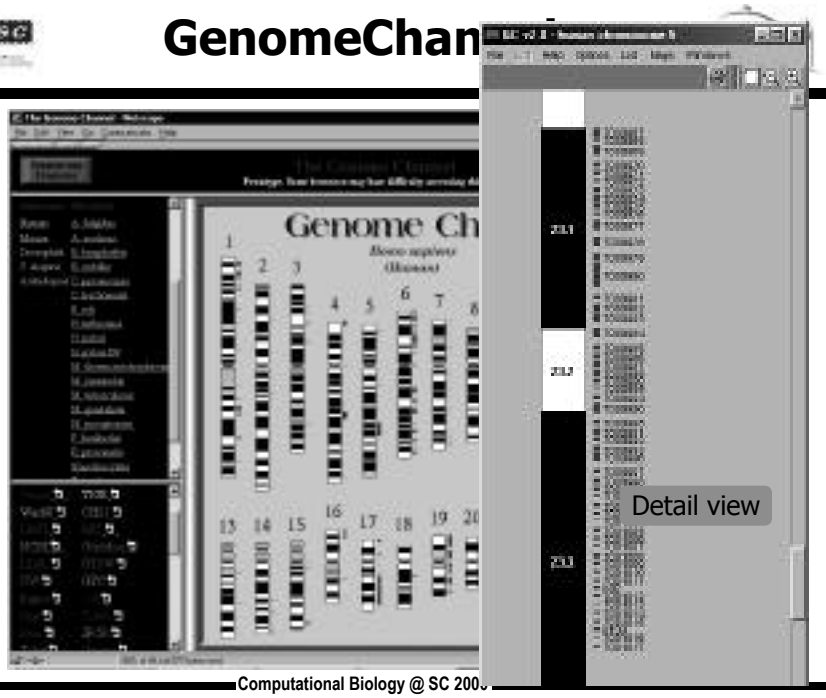
Annotation Pipeline



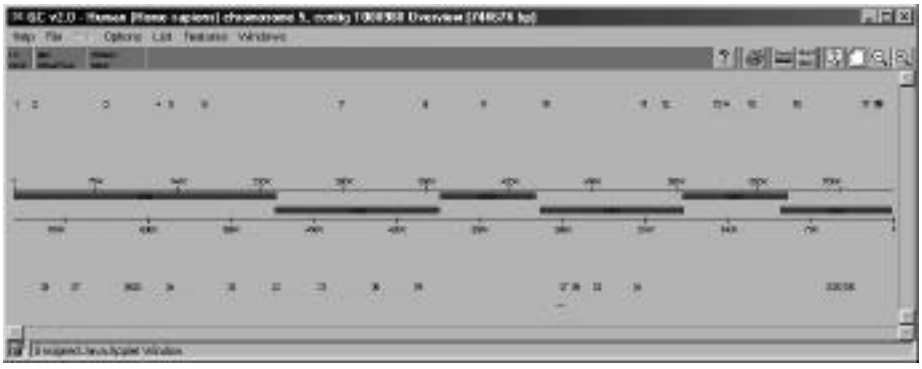
GenomeChannel



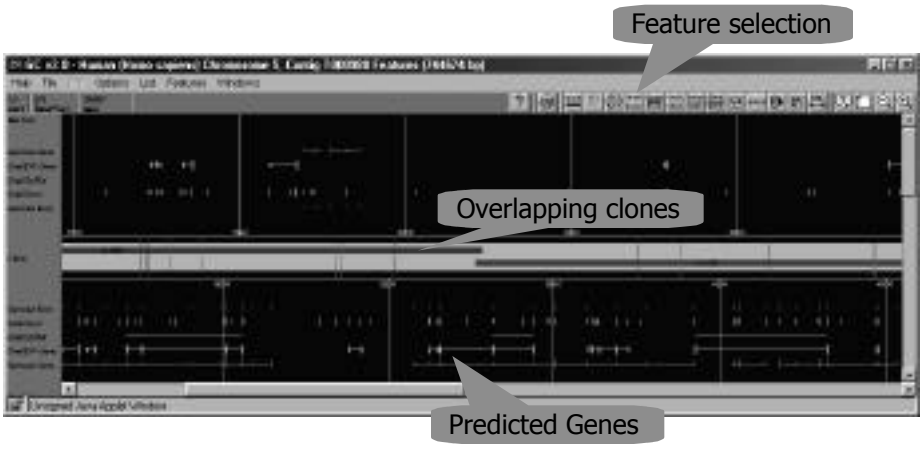
GenomeChan



A Contig Overview



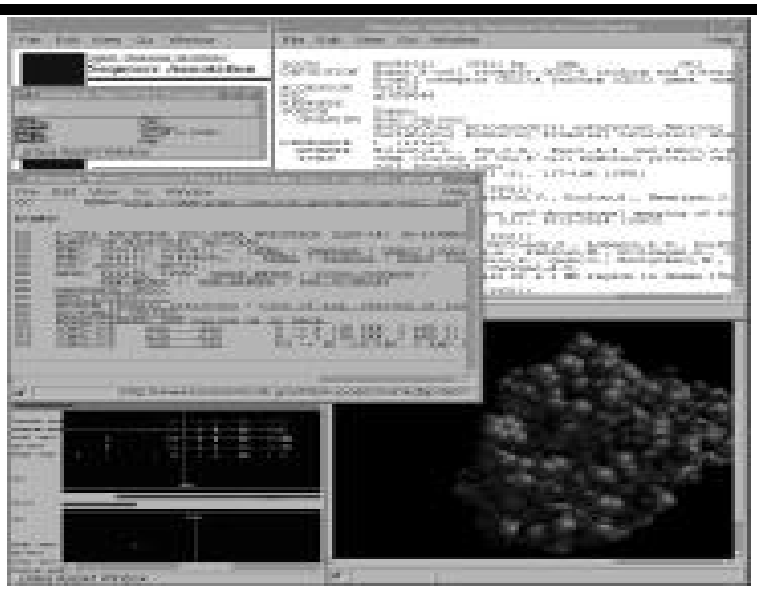
Feature Display



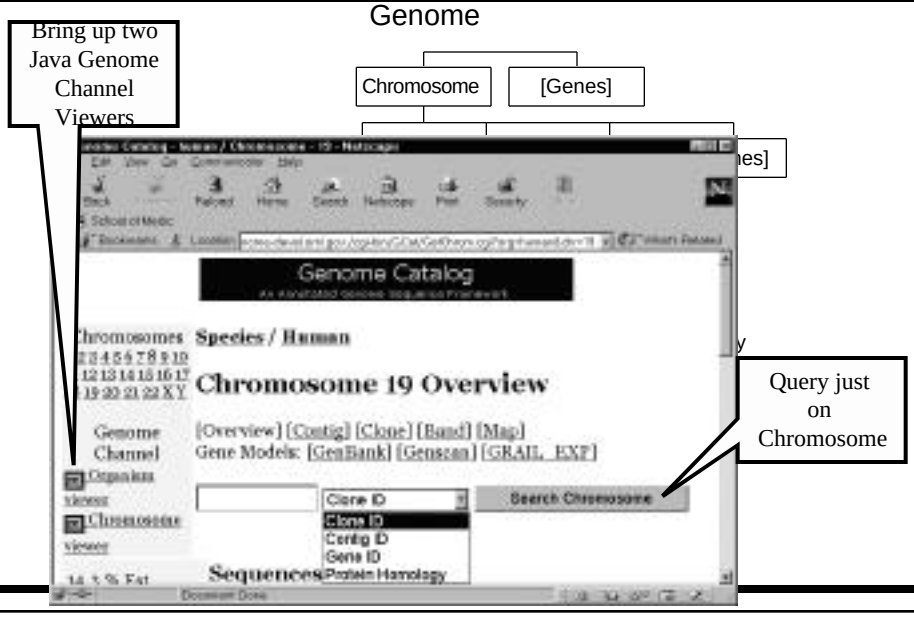
Gene Summary Report

BEAUTY - Gene Search Results

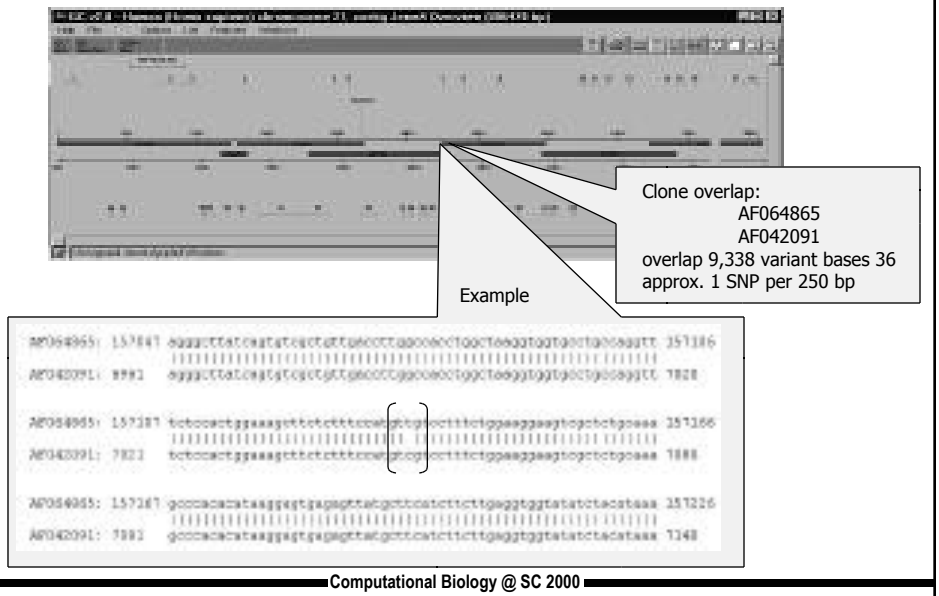
Reports and Links



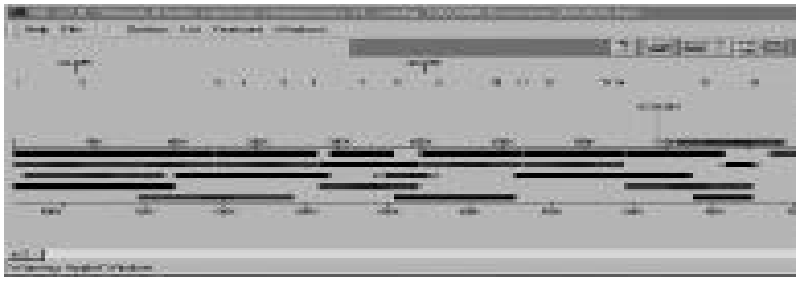
Navigate from human chromosome



SNP Mining from Clone Overlaps



SNP Mining from Clone Overlaps



Coverage includes clones from different sources
1 SNP per 250 bases
160,000 SNPs in 408 Mb dataset

What's supercomputing got to do with it?

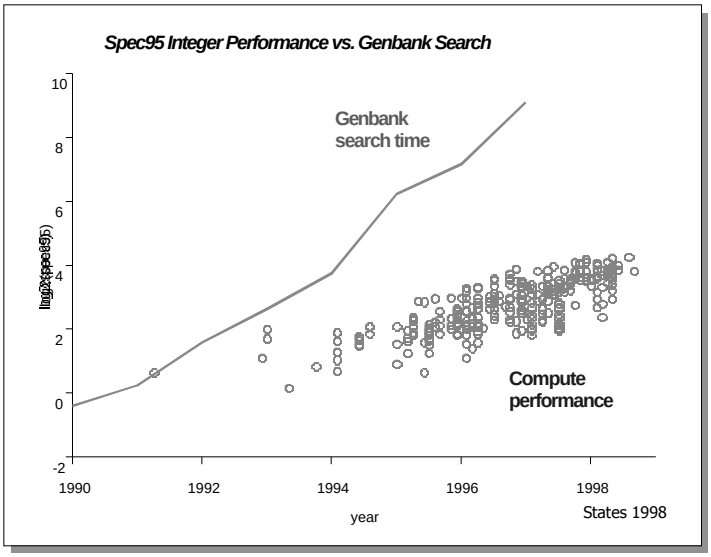
- **Complexity of the information**
- **Amount of data**
- **Most applications are trivially parallel**

Layers of Information

The same base sequence contains many layered instructions!

- **Chromosome structure and function**
  - **Telomers, centromers**
- **Gene Regulatory information**
  - **Enhancers, promoters, ...**
- **Instructions for gene structure**
- **Instructions for protein**
- **Instructions for protein post-processing and localization**

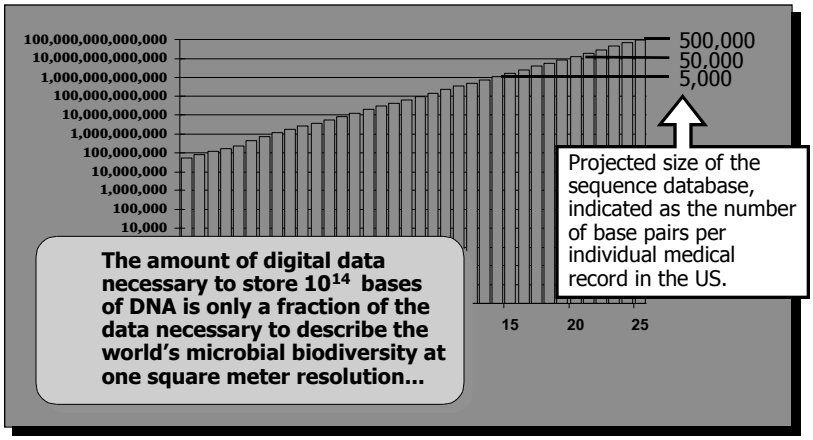
Moore's Law and Genomics



CPU Requirements

- **Current annotation**
  - **250 Mbases DNA yield ~125 Gbytes of data**
  - **It takes ~ 7.5 days on 20 workstations ~3,600nhr**
- **Celera Sequencing**
  - **Assembly of 1.7 Million reads in 25 hrs**
  - **Annotation 8-10 Mbases per months with 6 FTE**
  - **Assembly of Human Genome: expected ~ 3 months**

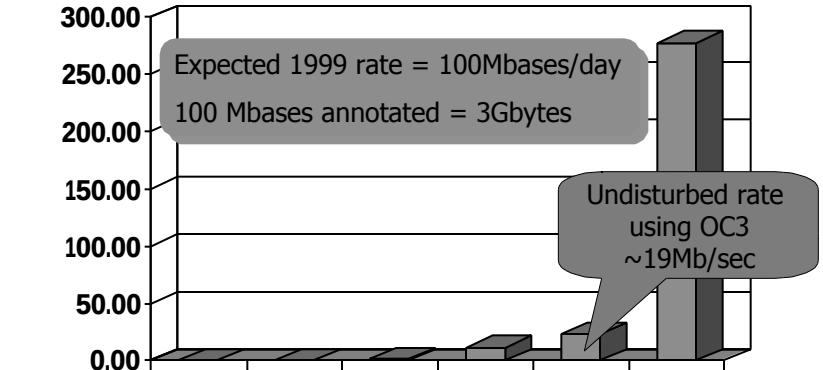
Projected Base Pairs



Sequence Assembly

- **Complexity**
  - **Adding a day's read of 100 Mb to a billion base pairs of contig would require 100 Pops operations**
  - **A 1 Tops machine would take about one day to process 100 Mbases**

Data Transfer

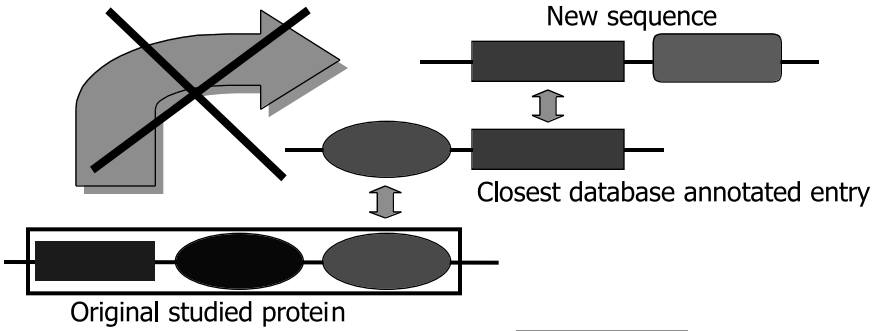


	year	month	week	day	12 hours	1 hour
Mbytes/sec	0.03	0.39	1.65	11.60	23.10	277.70

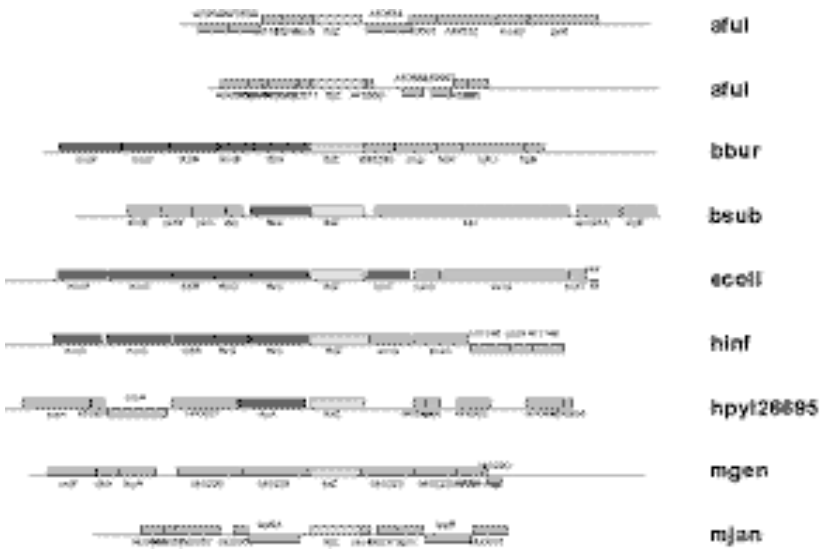
Challenges

- **Discovering new biology**
- **Lack of software integration**
- **Beginning to build high-performance applications**
- **Shortage of personnel**

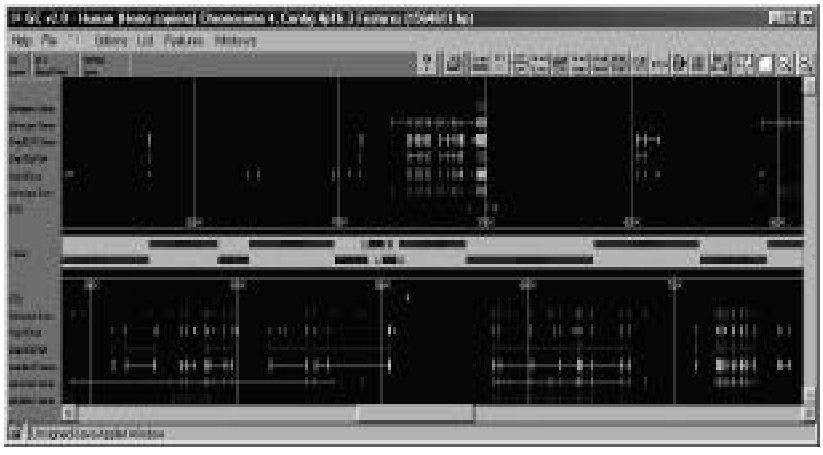
Inherited Annotation Problems in Multi-Domain Proteins



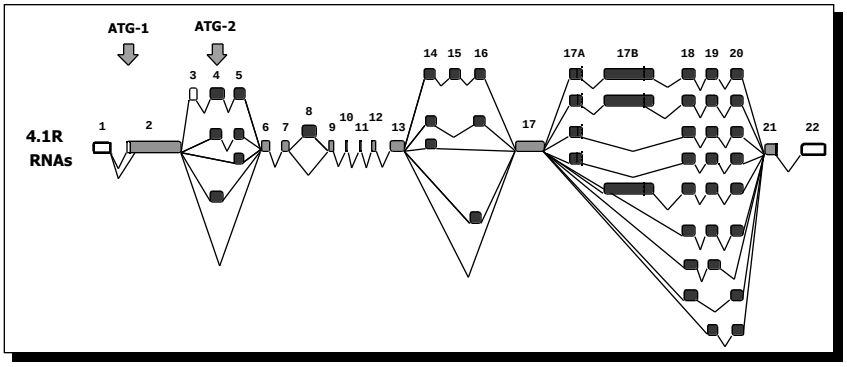
Comparative Genome Analysis



Alternatively Spliced ?



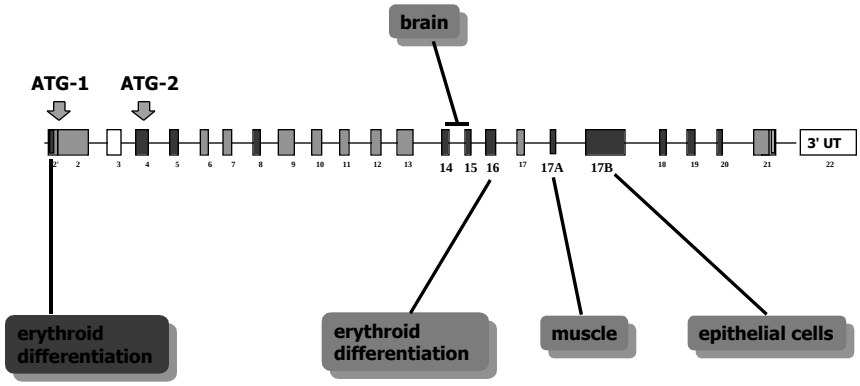
One Gene - Many Proteins



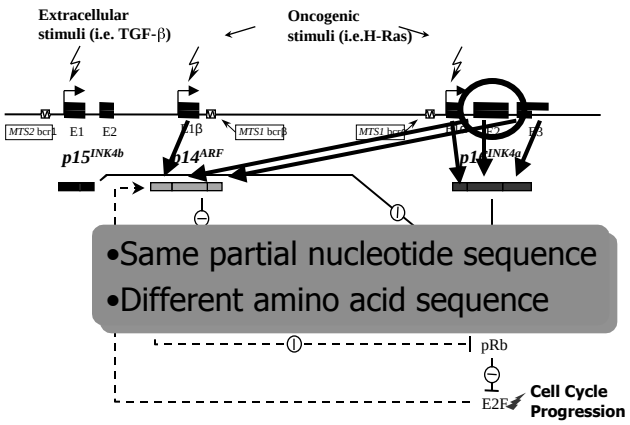
Conboy 1998

As many as 30% of human genes, in particular structural genes, may be alternatively spliced.

One Gene - Many Proteins



9p21 Gene Cluster is a Nexus of the Rb and p53 Pathways



Credits

- **NERSC / LBNL**
  - John Conboy
  - Donn Davy
  - Inna Dubchak
  - Sylvia Spengler
  - Denise Wolf
  - Eric P. Xing
  - Manfred Zorn
- **ORNL**
  - Ed Uberbacher
  - Richard Mural
  - Phil LoCascio
  - Sergey Petrov
  - Manesh Shah
  - Morey Parang